Tree Based Gibbs Sampling for Hierarchical Topic Model

Orbel 26 Mai 2023

What is topic modelling

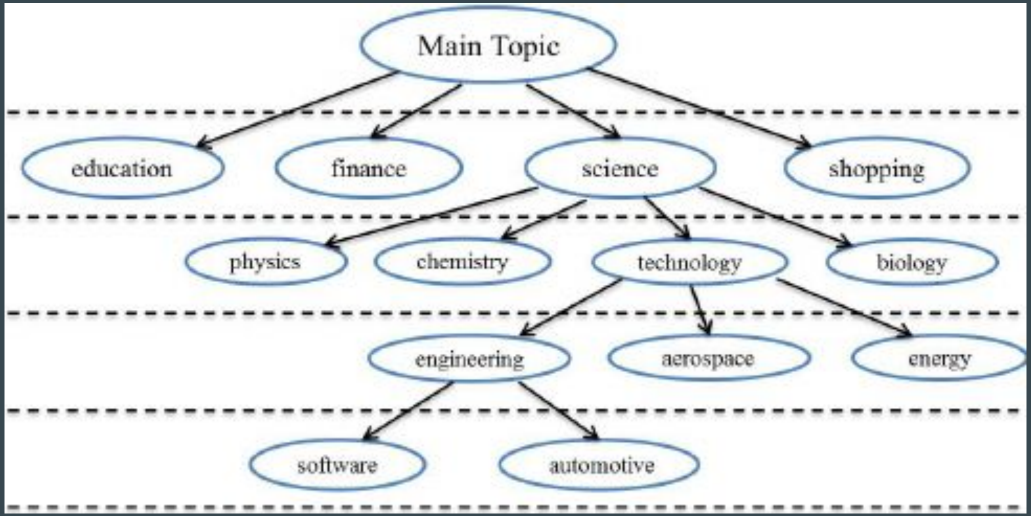
- Unsupervised text mining method
- To discover sets of co-occurring words inside documents (topics)

Applications of topic models

- To explore a dataset and quickly understand its content
  - Acts as a dimension reduction method
  - Helps in the data cleaning process
- As pre-processing to cluster similar documents
- Trend analysis (if applied to news articles)
- Text summarization (if applied to paragraphs)
- Customer segmentation (if applied to reviews/comments)
- Text classification based on document topic-distribution
- And countless other applications

A small genealogy of topic models

- LDA
  - One of the first and the most popular
  - Extracts k topics for a ***given*** k
  - Trained with Gibbs sampling
- HDP
  - No need to choose a k
  - Trained with Gibbs sampling
- nHDP
  - Models topic hierarchy
  - Trained with SVI
- HTMOT
  - Our model
  - Integrates temporal information
  - Trained with tree based Gibbs sampling



Training topic models

- Two methods
    - Stochastic Variational Inference
        - Fast
        - Not asymptotically exact
            - Good for large dataset
        - Hard to integrate distribution with no conjugate prior
    - Gibbs sampling
        - Slow (especially for complex models)
            - Bad for large dataset
        - Simple to integrate distribution with no conjugate prior
        - Asymptotically exact
            - Good for small dataset

Training topic models

- In HTMOT
  - We want to accurately extract small sub-topics
    - Represents small subset of the data
  - We want to model temporality as well
    - No conjugate prior for the beta distribution
- Gibbs sampling is best
  - But slower for large dataset
  - How can we improve the speed of the Gibbs sampling procedure?

Tree-based Gibbs sampling

Classic Gibbs sampling

- Six distributions to estimate
  - Topic-time distribution
  - Topic-word distribution
  - Document-topic distribution
  - Corpus-topic distribution
  - Topic hierarchy distribution
  - Topic word assignment distribution

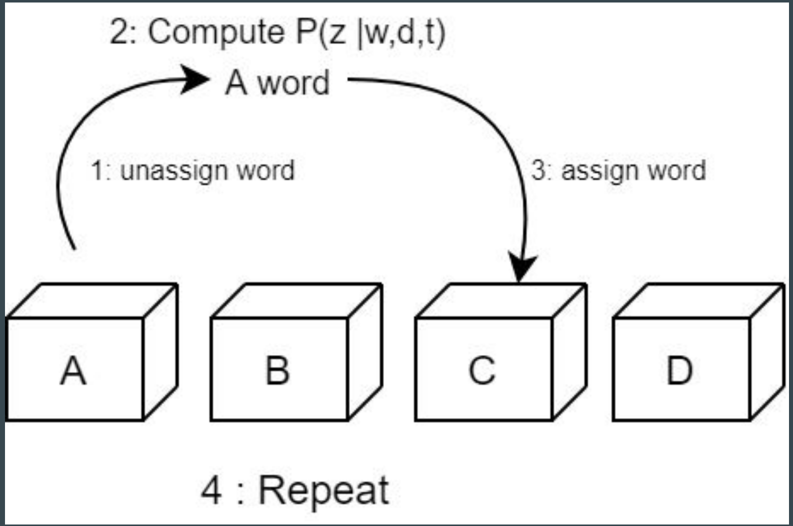
Algorithm 1 Traditional Gibbs sampling

procedure GIBBS(*corpus*)
    for N iterations **do**
        for each *document* in *corpus* **do**
            for each *word* in *document* **do**
                Sample word-topic assignment $P(z|w,d,t,B,D,T,C,H)$
                Sample topic-word $P(B|w,d,t,z,D,T,C,H)$
                Sample document-topic $P(D|w,d,t,B,z,T,C,H)$
                Estimate time-topic $P(T|w,d,t,B,D,z,C,H)$
                Sample corpus-topic $P(C|w,d,t,B,D,T,z,H)$
                Sample hierarchy-topic $P(H|w,d,t,B,D,T,C,z)$
            end for
        end for
    end for
    Return solution : (z,B,D,T,C,H)
end procedure

- Classic Gibbs sampling
  - Sample from all six distributions iteratively
    - Complexity is linear with respect to # of variables in the model
  - It’s really slow to begin with
    - Worse for complex models

Bin-based Gibbs sampling

- Simple solution :
  - Only sample from the
    - Topic word assignment distribution
  - Let a data structure do the rest



"Bin counting" Gibbs sampling (z points to one of the box)

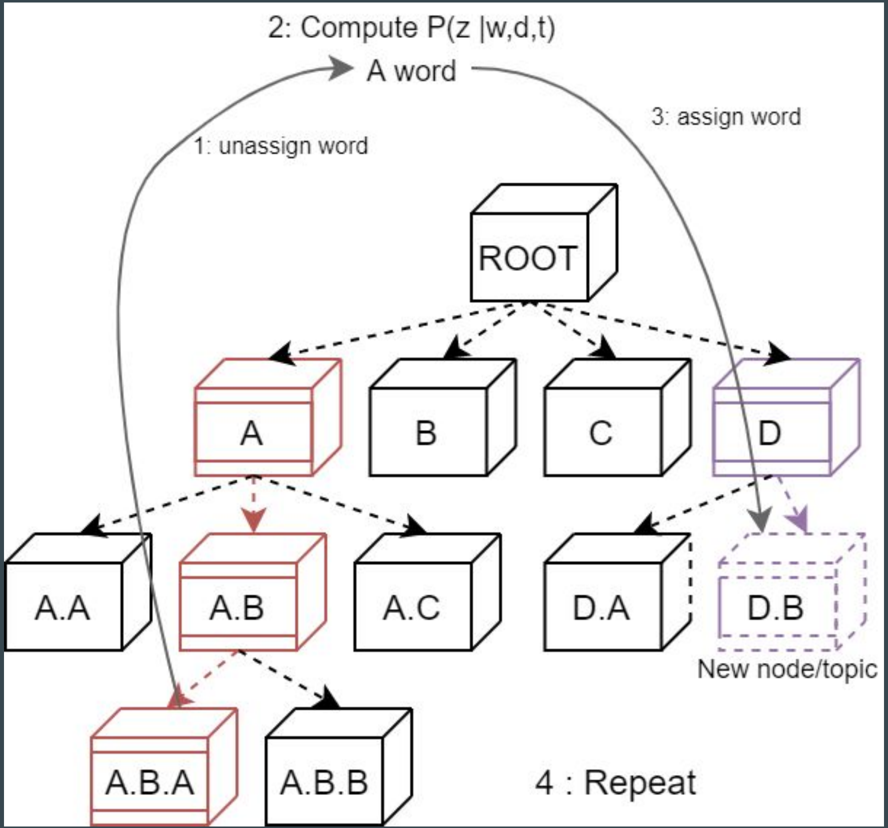
- Basic exemple
  - Assign every word in corpus randomly to bins
  - For each word sample the Topic word assignment distribution
    - Condition on all other variables
    - Estimated from bin content
  - Re-assign the word to the new chosen bin
  - Over time the bins are sorted into topics

Bin-based Gibbs sampling

- When & where are the other distributions estimated?
  - By the bins themselves
    - Each bin is a topic, its content defines a topic-word distribution
    - Each bin is a topic, together the bins define a corpus-topic distribution
    - Each word is associated with a document
      - If we only count words associated to a document
      - Together the bins define a document-topic distribution
    - …
- => All other distributions are approximated by moving words around

Tree-based Gibbs sampling

- Infinite Dirichlet Trees
  - Same idea, but nested
  - Each time a word is assigned to a bin
  - We draw from a bernoulli to decide to go deeper
  - If yes, we repeat the same process on sub-topics
  - + added small probability to draw a new bin
    - To decide # of topics during training
- Words assignments are now random path in the Tree
- One tree for the corpus
- One tree for each documents
  - Mutually exclusive subset of corpus tree



Sampling from P(z|w,d,t)

- When drawing a topic assignment for a word,
  - we either draw from the local tree with some probability
  - or we draw from the corpus tree with some other probability
  - or else we create a new topic.
- Then, we draw from a Bernoulli to decide if we go deeper or not.
  - If we do go deeper we repeat the same process until we eventually stop.
- We repeat this process for each word of each document until convergence

$$z|w,d = \begin{cases} \sum_k \frac{(1+BetaPDF(\rho_k^1*\Delta_j,\rho_k^2*\Delta_j))*(A(k|d)+\epsilon)*(A(k|w)+\phi)*\delta_k}{(A(k)+(\phi*V))*m_d} \\ \sum_k \frac{(1+BetaPDF(\rho_k^1*\Delta_j,\rho_k^2*\Delta_j))*(A(k|w)+\phi)*\delta_k}{m_w} \\ new \end{cases}$$

j

Selecting a sibling at depth j

$$p = \frac{\boxed{P}+\boxed{\theta_1}}{\boxed{N}+\boxed{\theta_1}+\boxed{\theta_2}+\boxed{C}+\boxed{P}}$$

$$\boxed{P} = \frac{(1+BetaPDF(\rho_k^1,\rho_k^2))*(A^*(k|w)+\phi)*(A^*(k|d)+\epsilon)}{A^*(k)+(\phi*V)}$$

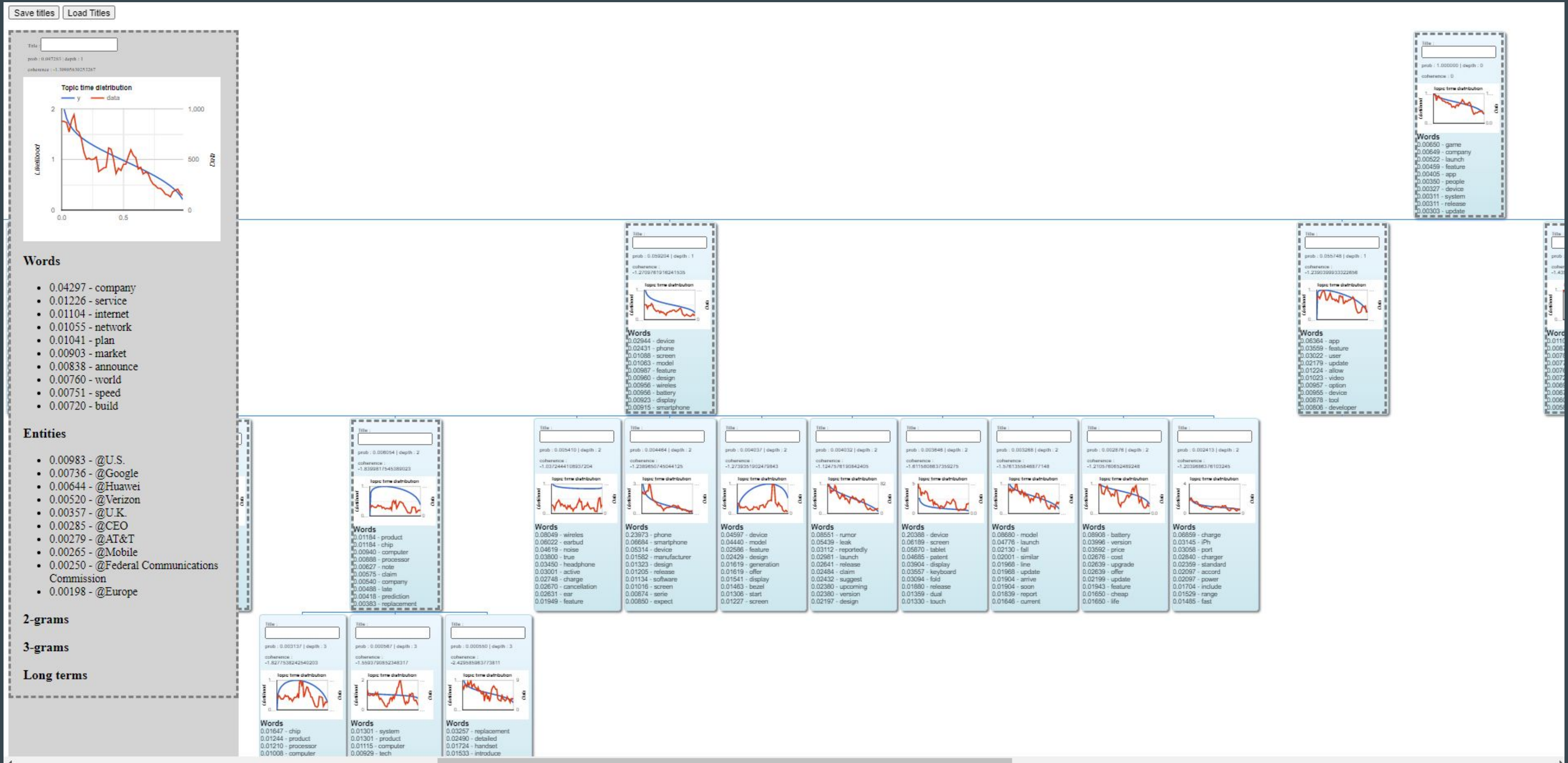
$$\boxed{N} = \frac{\phi*\epsilon}{\phi*V}$$

$$\boxed{C} = \sum_i \frac{(1+BetaPDF(\rho_i^1,\rho_i^2))*(A(i|w)+\phi)*(A(i|d)+\epsilon)}{A(i)+(\phi*V)}$$

Deciding whether to go deeper

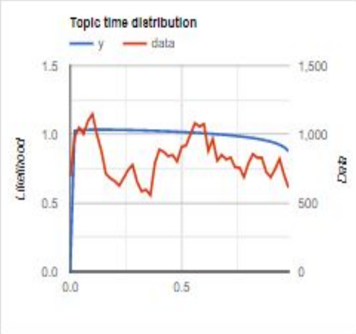
Results

Results : interface



Results : examples

- Space
  - Astronomy
  - Astronauts
- Topic are coherent
  - Content, Hierarchy, & time

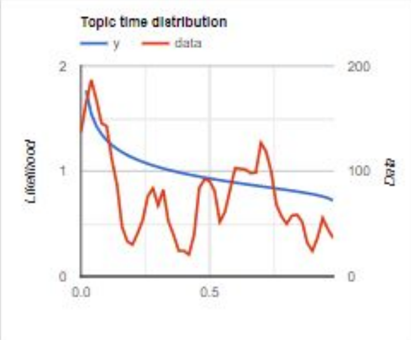


Words

- 0.02499 - launch
- 0.02182 - space
- 0.02003 - mission
- 0.01268 - planet
- 0.01252 - astronaut
- 0.01113 - rocket
- 0.00963 - test
- 0.00939 - spacecraft
- 0.00914 - satellite
- 0.00845 - orbit

Entities

- 0.02839 - @NASA
- 0.01395 - @Earth
- 0.00894 - @International Space Station
- 0.00843 - @Mar
- 0.00688 - @SpaceX
- 0.00525 - @Dragon
- 0.00457 - @Crew
- 0.00291 - @Starlink
- 0.00254 - @Florida
- 0.00237 - @European Space Agency

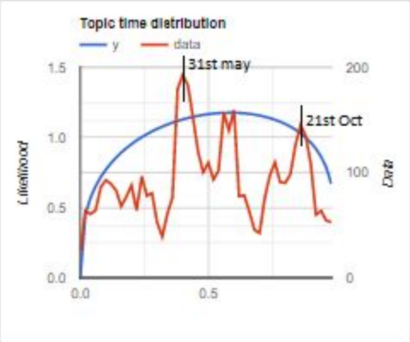


Words

- 0.05180 - galaxy
- 0.03595 - black
- 0.03462 - hole
- 0.02630 - star
- 0.02537 - image
- 0.02471 - telescope
- 0.02181 - light
- 0.01533 - ga
- 0.01335 - astronomer
- 0.01163 - observe

Entities

- 0.01097 - @Hubble
- 0.00846 - @ESA
- 0.00740 - @Way
- 0.00687 - @Milky
- 0.00595 - @Spitzer
- 0.00581 - @Telescope
- 0.00449 - @Hubble Space Telescope
- 0.00383 - @Observatory
- 0.00330 - @NGC
- 0.00291 - @Southern



Words

- 0.07890 - astronaut
- 0.05679 - space
- 0.04230 - crew
- 0.03757 - station
- 0.02684 - mission
- 0.02222 - spacecraft
- 0.01900 - capsule
- 0.01106 - flight
- 0.00913 - launch
- 0.00913 - return

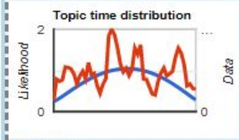
Entities

- 0.06076 - @International Space Station
- 0.05335 - @NASA
- 0.03328 - @Dragon
- 0.02984 - @Crew
- 0.01707 - @SpaceX
- 0.01578 - @Earth
- 0.00730 - @American
- 0.00719 - @Behnken
- 0.00709 - @Bob Behnken
- 0.00644 - @Hurley

Title : Astraunauts

key : 7062 | prob : 0.011791 | depth : 2 | KL GM : 1.403641

coherence : -0.42941789743321207



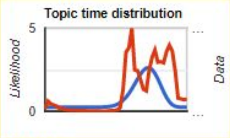
Words

0.08280 - astronaut
0.04829 - space
0.04122 - crew
0.03738 - station
0.03425 - mission
0.02110 - spacecraft
0.01914 - capsule
0.01055 - program
0.00948 - flight
0.00903 - return

Title : Return of astronauts

key : 828 | prob : 0.003625 | depth : 3 | KL GM : 1.031186

coherence : -0.4308320138002019



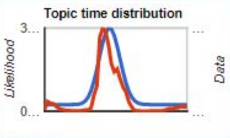
Words

0.08755 - astronaut
0.05090 - space
0.04799 - station
0.04538 - crew
0.04014 - mission
0.02414 - spacecraft
0.01949 - capsule
0.01251 - flight
0.01076 - return
0.00785 - operational

Title : SpaceX first crewed flight

key : 6939 | prob : 0.002911 | depth : 3 | KL GM : 1.031731

coherence : -0.4308320138002019



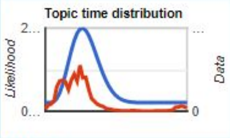
Words

0.08656 - astronaut
0.04962 - space
0.03405 - station
0.03368 - crew
0.02970 - mission
0.02209 - capsule
0.02137 - spacewalk
0.02028 - spacecraft
0.01231 - program
0.00978 - shuttle

Title : ISS 3 new astronauts

key : 5023 | prob : 0.002389 | depth : 3 | KL GM : 1.027453

coherence : -0.3990310530626589



Words

0.11342 - astronaut
0.06311 - crew
0.05781 - space
0.04722 - mission
0.03883 - station
0.02383 - spacecraft
0.01898 - capsule
0.01721 - program
0.01324 - cargo
0.01015 - carry

Any questions?